conventional brain scans are not specific for the disease. It is difficult to interpret enhancing lesions on magnetic resonance imaging in multiple sclerosis. Contrast enhancement in images indicates local breakdown of the blood-brain barrier, presumably owing to focal inflammation in multiple sclerosis, but contrast enhancement in the white matter after stress or hypoxia is due to inflammation. Surrogate markers based on imaging results used as outcome measures in multiple sclerosis trials do not mirror the clinical course of the disease. Although neurodegeneration is probably the most important cause of fixed and progressive disability in multiple sclerosis,[18] imaging surrogates for neuroaxonal loss have not been validated for predicting future disability.[19]

Experimental allergic encephalomyelitis is not a suitable animal model for testing treatments for multiple sclerosis and it is time to explore alternative experimental and therapeutic approaches.[14 20] Clinical research is needed to reveal the biological variables that can distinguish relapsing progressive disease from relatively benign disease. A successful treatment should delay progressive tissue loss irrespective of relapse rates and clinical phenotype. Unusually for a neurological disease, the therapeutic time window for intervention is wide in multiple sclerosis, so that research on neuroprotective strategies should be a priority. Short term solutions for a chronic disease like multiple sclerosis are not likely to be effective, and PML resulting from treatment with natalizumab should be taken as a signal to change the way we treat this disease.

Contributors and sources: AC has a research interest in multiple sclerosis and therapeutics in neurology. This article was developed from the discussions with colleagues on new treatment trials in multiple sclerosis. AC is the sole contributor and guarantor of the article.

Competing interests: None declared.

1 Van Assache G, Ranst MV, Sciot R, Dubois B, Vermeire S, Noman M, et al. Progressive multifocal leukoencephalopathy after natalizumab therapy for Crohn's disease. *N Engl J Med* 2005;353:362-8.
2 Keinschimdt-DeMasters BK, Tyler KL. Progressive multifocal leukoencephalopathy complicating treatment with natalizumab and interferon beta-1a for multiple sclerosis. *N Engl J Med* 2005;353:369-74.
3 Langer-Gould A, Atlas SW, Bollen AW, Pelletier D. Progressive multifocal leukoencephalopathy in a patient treated with natalizumab. *N Engl J Med* 2005;353:375-81.
4 Rice GPA, Hartung HP, Calabresi PA. Anti-alpha4 integrin therapy for multiple sclerosis: mechanism and rationale. *Neurology* 2005;64:1336-42.
5 US Food and Drug Administration. Center for Drug Evaluation and Research. *Tysabri (Natalizumab).* www.fda.gov/cder/foi/nda/2004/125104s000_Natalizumab.htm.
6 US Food and Drug Administration. FDA issues public health advisory on Tysabri, a new drug for MS. www.fda.gov/bbs/topics/news/2005/NEW01158.html (accessed 30 Jun 2005).
7 Chaudhuri A, Behan PO. Natalizumab for multiple sclerosis. *N Engl J Med* 2003;348:1598-9.
8 Tubridy N, Behan PO, Capildeo R, Chaudhuri A, Forbes R, Hawkins CP, et al. The effect of anti-alpha-4-integrin antibody on brain lesion activity in MS. *Neurology* 1999;53:466-72.
9 Miller DH, Khan OA, Sheremata WA, Blumhardt LD, Rice GPA, Libonati MA, et al. A controlled trial of natalizumab for relapsing multiple sclerosis. *N Engl J Med* 2003;348:15-23.
10 O'Connor PW, Goodman A, Willmer-Hulme AJ, Libonati MA, Metz L, Murray RS, et al. Randomised multicenter trial of natalizumab in acute MS relapses. Clinical and MRI effects. *Neurology* 2004;62:2038-43.
11 Kleinke JD, Gottlieb S. Is the FDA approving drugs too fast? *BMJ* 1998;317:899-900.
12 Bjursten M, Bland PW, Willen R, Hornquist EH. Long-term treatment with alpha-4 integrin antibodies aggravates colitis in G-αi2-deficient mice. *Eur J Immunol* 2005:2274-83.
13 Berger JR, Koralnik IJ. Progressive multifocal leukoencephalopathy and natalizumab—unforeseen consequences. *N Engl J Med* 2005;353:414-6.
14 Chaudhuri A, Behan PO. Treatment of multiple sclerosis beyond the NICE guideline. *QJM* 2005;98:373-8.
15 Confraveux C, Vukusic S, Adeleine P. Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. *Brain* 2003;126:770-82.
16 Eriksson M, Andersen O, Runmaker B. Long term follow up of patients with clinically isolated syndromes, relapsing-remitting and secondary progressive multiple sclerosis. *Mult Scler* 2003;9:260-74.
17 *ICH harmonised tripartite guideline for good clinical practice.* www.ich.org/LOB/media/MEDIA482.pdf (accessed 16 Jan 2006).
18 Chaudhuri A. Interferon beta, progressive MS, and brain atrophy. *Lancet Neurol* 2005;4:208-9.
19 Miller DH. Biomarkers and surrogate outcomes in neurodegenerative diseases: lessons from multiple sclerosis. *NeuroRx* 2004;1:284-94.
20 Chaudhuri A, Behan PO. Multiple sclerosis: looking beyond autoimmunity. *J R Soc Med* 2005; 98:303-6.

(Accepted 9 October 2005)

Health policy

Have targets improved performance in the English NHS?

Gwyn Bevan, Christopher Hood

The star rating system for NHS trusts seems to have improved performance, but we still don't know how genuine the improvements are or the costs to other services

Annual performance ratings have been published for NHS trusts in England since 2001, and the fifth and final set was published in July 2005.[1–6] This process of naming and shaming gave each trust a rating from zero to three stars. Trusts that failed against a small number of key targets were at risk of being zero rated and their chief executives at risk of losing their job; trusts that performed well achieved three stars and were eligible for benefits from "earned autonomy."[7] Although the government has abandoned the star ratings, targets are likely to remain. We consider reported improvements in performance against key targets, problems of the system, and what ought to happen in the future.

Reported improvements in performance

We compared data on performance in England before and after the star rating system for three key targets. When data were available we also compared English data with that of other UK countries that did not adopt the star system.

Accident and emergency departments

The key target for accident and emergency departments was the percentage of patients to be seen within four hours. From March 2003, the target was 90%,[3 5]

Department of Operational Research, London School of Economics and Political Science, London WC2A 2AE
Gwyn Bevan *professor of management science*

continued over

BMJ 2006;332:419–22

References w1-w12 and sources of data are on bmj.com

All Souls College, University of Oxford, Oxford OX1 4AL
Christopher Hood
Gladstone professor of government

Correspondence to:
G Bevan
R.G.Bevan@lse.ac.uk

and from January 2005 this increased to 98%.[6] The National Audit Office reported that in England, in 2002, 23% of patients spent over four hours in accident and emergency, but in the three months from April to June 2004 only 5.3% stayed that long [w1]; this increased patient satisfaction and was achieved despite increasing use of emergency services.

Ambulance category A calls

England has had a target for category A calls (life threatening emergencies) since 1996, before star ratings were applied to ambulance trusts. The target was that that at least 75% of calls be met within 8 minutes[w2]; this became a key target from the end of 2002, when star ratings applied to ambulance trusts.[2 3 5 6] About 30 trusts have provided ambulance services. Comparable data are available for 17 trusts for the two years before, and the four years during which, star ratings applied.[w3] For the year ending in March 2000, only one trust had response rates above 75% and two trusts had rates lower than 40%. Reported performance improved greatly after ambulance trusts were star rated. For the year ending in March 2005, 14 trusts exceeded the target and the worst performer achieved 71%.

The Welsh Ambulance Service NHS Trust also had the target of responding to 75% of category A calls within 8 minutes by the end of 2001.[w4] Response rates, however, remained at about 50% between 2001 and 2004.[w5]

First elective hospital admission

The key target for first elective hospital admission was the maximum wait: this was 18 months by the end of March 2001,[1] 15 months by 2002,[2] 12 months by 2003,[3] and 9 months by 2004.[5 6] The numbers of patients waiting more than 12 and 9 months in England at the end of March 1998 were reported to be 67 000 and 185 000, but by the end of March 2005, only 24 were reported to be waiting more than 12 months and 41 more than 9 months.[w6]

Table 1 gives the percentages of patients waiting for more than six and 12 months at the end of March from 1999 to 2005 for England, Wales, and Northern Ireland. From 2001 to 2003, reported performance improved in England but deteriorated in Wales and Northern Ireland. After that, however, reported performance improved in all countries, dramatically in Wales and Northern Ireland. This suggests that the policy of naming and shaming in England put pressure on the NHS in the other countries.

Table 1 Percentages of patients on NHS hospital waiting lists waiting longer than six or 12 months, 1999-2005

	1999	2000	2001	2002	2003	2004	2005
% waiting >12 months							
England	4.4	4.7	4.2	2.1	0	0	0
Wales	11.2	14.2	13.8	14.3	15.9	11.3	1.3
Northern Ireland	17.9	20.0	21.8	24.9	22.0	14.7	8.5
% waiting >6 months							
England	26.1	25.8	24.4	23.3	19.4	8.9	5.0
Wales	NA	NA	34.0	37.0	37.0	35.2	24.9
Northern Ireland	36.7	39.1	41.4	44.1	40.0	34.1	28.1

NA=data not available.

Problems with targets

Star ratings have been criticised for their similarities to the target regime of the former Soviet Union, although NHS managers were threatened with loss of their jobs rather than their life or liberty.[7 8] The Soviet target regime seemed to produce substantial improvements in the 1930s but was recognised to have serious problems from the 1950s and collapsed in the 1990s.[9] In May 2005, during the British general election campaign, the prime minister was apparently nonplussed by a complaint made during a televised question session that pressure to meet the key target that 100% of patients be offered an appointment to see a general practitioner within two working days[6] had meant that many general practices refused to book any appointments more than two days in advance.[w7] A survey of patients found that 30% reported that their general practice did not allow them to make a doctor's appointment three or more working days in advance.[w8] Many saw the perverse outcome of a key target that was intended to improve access to general practitioners as a reason for abandoning the system of targets and star ratings.

Regulation by targets assumes that priorities can be targeted, the part that is measured can stand for the whole, and what is omitted does not matter. But most indicators of healthcare performance are "tin openers rather than dials . . . they do not give answers but prompt investigation and inquiry, and by themselves provide an incomplete and inaccurate picture."[10] Hence, typically for defined priorities there will be a few good measures ("dials," such as waiting times); a larger group of imperfect measures ("tin openers," such as mortality), the use of which is liable to generate false positive and false negative results; and an even larger group for which no usable data are available (which applies to the clinical quality of much of health care[10]). This last group was the cause of the neglect of quality in the Soviet regime, which was widely claimed to be an endemic problem from Stalin to Gorbachev.[9]

The use of targets results in gaming,[7–9 11–13] which means that when reported performance meets the targets, neither government nor the public can distinguish between the following four outcomes:

- All is well; performance has been exactly as desired in all domains (whether measured or not)

Table 2 Evidence of gaming in response to three type of targets

Problem	Target		
	<4 hour wait in accident and emergency	Ambulance category A calls*	Maximum waiting times for first elective hospital admission
Poor performance in domains where performance not measured	Extra staff drafted in and operations cancelled for the period over which performance was measured[w9 w10]	Strong allegations that some ambulance trusts relocated depots from rural to urban areas hence achieving the target at the expense of a worse service in rural areas[15]	
Hitting the target and missing the point	Patients had to wait in ambulances outside the department until staff were confident of meeting the target[15]	Idiosyncrasies in the rules of classification led to some patients in urgent need being given a lower priority than less serious cases[15]	Patients may have been removed from waiting lists once they had been provided with a future date for an appointment, or given immediate appointments that they were not able to attend and then classed as refusing treatment, or had treatment inappropriately suspended[16]
Ambiguity in reporting of data or fabrication	The level reported to the Department of Health in 2004-5 was 96%, but an independent survey of patients reported only 77%[8]	Problems in the definition of category A calls (the proportion of logged calls varied by more than fivefold) and ambiguity in the time when the clock started.[12 13] A third of ambulance trusts had "corrected" response times to be less than 8 min[15]	Nine NHS trusts had "inappropriately" adjusted their waiting lists[w11]; three others had deliberately misreported waiting list information; and 19 trusts had reporting errors in at least one indicator[w12]

* Response within 8 minutes for 75% of calls.

- The organisation's performance has been as desired where performance was measured but at the expense of unacceptably poor performance in the domains where performance was not measured
- Although reported performance against targets seems to be fine, actions have been at variance with the substantive goals behind those targets (hitting the target and missing the point)
- Targets have not been met, but this has been concealed by ambiguity in the way data are reported or outright fabrication.

Table 2 presents evidence that these problems have occurred in the three key targets discussed above. Although we have no evidence of poor performance in other domains in response to the target for inpatient waiting times, this type of gaming was reported for the target for new outpatient waiting times. Ophthalmology services in Bristol met that target by cancelling and delaying follow-up outpatient appointments (which had no target) and, as a consequence, at least 25 patients were estimated to have lost their vision over two years.[13] The Audit Commission's last report based on spot checks of the quality of data in 55 trusts concluded that the scale of reporting errors identified did not undermine the reliability of overall trends reported nationally.[14] But questions remain over the extent to which improvements in targeted performance in the English NHS were undermined by other types of gaming and whether similar problems underlie the big reductions in long waiting times reported in Wales and Northern Ireland in 2004 and 2005.

What next?

Nobody would want to return to the NHS performance before the introduction of targets, with over 20% of patients spending more than four hours in accident and emergency and patients waiting more than 18 months for elective admission. And attempts to improve performance without the star system in Wales were criticised by the auditor general for Wales for having "provided neither strong incentives nor sanctions to improve waiting time performance" and were widely perceived to have rewarded organisations that failed to deliver on waiting times.[17] So how can we maximise the social benefits and minimise the costs of a regime of targets with sanctions?

We suggest two remedies. One, for which we have argued earlier,[18] is to introduce more uncertainty in the way that performance will be assessed and thus make some kinds of managerial gaming more difficult. A second is to remedy the continuing lack of coherent systematic auditing of performance data of the healthcare system in England. Despite the heavy regulatory burden from auditors and assessors of various kinds, if anything the audit hole is getting bigger. Current proposals for assessing performance seem to favour reliance on statistical data to assess the robustness of performance data[19] rather than regular visits by the Commission for Health Improvement, which uncovered gaming practices.[15 16] In addition, responsibility for auditing the quality of data in the English NHS has been transferred from the Audit Commission to the Healthcare Commission, which has no presence on the ground in NHS provider units.[14]

We need an independent body that approximates to the Office of Performance Data advocated by Robert Behn.[20] Such a body would investigate the genuineness of reported improvements in healthcare performance and whether improvements are achieved at the cost of what cannot be easily measured. Although these changes would not wholly eliminate the gaming problems associated with any regime of targets and terror, they could reduce them. The current combination of performance measures that are highly predictable to managers and an audit system that is poorly equipped to detect gaming systematically, risks losing credibility and the prospect of even more awkward questions being asked in the next general election campaign.

Summary points

The star rating system for English NHS trusts has improved reported performance on key targets

The effect on services excluded from star ratings is unclear

In some cases data have been manipulated to achieve targets

Systems need to be put in place to minimise gaming to meet targets and ensure targets are not causing unwanted effects elsewhere

We thank those who helped us identify comparable statistics with England for Wales and Northern Ireland and explained that no such statistics are available for Scotland. Also thanks to Olly Bevan for assembling the statistical material.

Contributors and sources: The evidence and ideas for this paper come from GB's involvement in the development of NHS star ratings in England and CH's extensive research into regulation by governments in various sectors. The article is based on numerous presentations by both authors. GB did the analysis of the impact and evidence of gaming and wrote the first draft. The concepts underlying the paper were developed jointly. CH contributed to revisions of the paper.

Competing interests: GB was director of the office for information on healthcare performance at the Commission for Health Improvement until September 2003.

1 Department of Health. *NHS performance ratings acute trusts 2000/01*. London: DoH, 2001. www.dh.gov.uk (search for: 25290).
2 Department of Health. *NHS performance ratings acute trusts, specialist trusts, ambulance trusts, mental health trusts 2001/02*. London: DoH, 2002. www.dh.gov.uk (search for: 28859).
3 Commission for Health Improvement. *NHS performance ratings. Acute trusts, specialist trusts, ambulance trusts 2002/03*. London: Stationery Office, 2003. http://ratings2003.healthcarecommission.org.uk/ratings/ (accessed 28 Sep 2005).
4 Commission for Health Improvement. *NHS performance ratings. Primary care trusts, mental health trusts, learning disability trusts 2002/03*. London: Stationery Office, 2003. http://ratings2003.healthcarecommission.org.uk/ratings/ (accessed 28 Sep 2005).
5 Healthcare Commission. *Performance rating*. London: Stationery Office, 2004. http://ratings2004.healthcarecommission.org.uk/ (accessed 28 Sep 2005).
6 Healthcare Commission. *NHS performance ratings 2004/2005*. London: Healthcare Commission, 2005. http://ratings2005.healthcarecommission.org.uk/ (accessed 28 Sep 2005).
7 Bevan G, Hood C. What's measured is what matters: targets and gaming in the English public health care system. *Public Admin* (in press).
8 Statistics and politics in Britain. *Economist* 2005 Mar 23.
9 Berliner JS. *Soviet industry from Stalin to Gorbachev*. Aldershot: Edward Elgar, 1988.
10 Carter N, Klein R, Day P. *How organisations measure success. The use of performance indicators in government*. London: Routledge, 1995.
11 Kornai J. *Overcentralisation in economic administration*. Oxford: Oxford University Press, 1994.
12 Bird SM, Cox D, Farewell VT, Goldstein H, Holt T, Smith PC. Performance indicators: good, bad, and ugly. *J R Stat Soc A* 2005;168: 1–27.
13 Public Administration Select Committee. *Fifth report on target? Government by measurement*. London: Stationery Office, 2003. www.publications.parliament.uk/pa/cm200203/cmselect/cmpubadm/62/62.pdf (accessed 28 Sep 2005).
14 Audit Commission. *Information and data quality in the NHS*. London: Audit Commission, 2004. www.audit-commission.gov.uk/reports/NATIONAL-REPORT.asp?CategoryID=&ProdID=4D598AF6-3894-401d-AA48-1076125DA38D (accessed 28 Sep 2005).
15 Commission for Health Improvement. *What CHI has found in ambulance trusts*. London: Stationery Office, 2003. www.healthcarecommission.org.uk/NationalFindings/NationalThemedReports/Ambulance/fs/en (accessed 28 Sep 2005).
16 Commission for Health Improvement. *What CHI has found in acute services*. London: Stationery Office, 2004. www.healthcarecommission.org.uk/NationalFindings/NationalThemedReports/AcuteAndSpecialist/fs/en (accessed 28 Sep 2005).
17 Auditor General for Wales. *NHS waiting times in Wales. Vol 2. Tackling the problem*. Cardiff: Stationery Office, 2005. www.wao.gov.uk/assets/englishdocuments/NHS_Waiting_Times_Vol_2_agw_2005.pdf (accessed 28 Sep 2005).
18 Bevan G, Hood C. Targets, inspections, and transparency. *BMJ* 2004;328:598.
19 Healthcare Commission. *The annual health check*. London: Healthcare Commission, 2005. www.healthcarecommission.org.uk/InformationForServiceProviders/AnnualHealthCheck/fs/en?CONTENT_ID=4017483&chk=ub2qrx (accessed 28 Sep 2005)
20 Behn R. *Rethinking democratic accountability*. Washington, DC: Brookings Institution, 2001.

(Accepted 17 November 2005)

An unusual ending to an anatomy lesson

I work in Phnom Penh, in a community health project. Despite thinking that I am well adjusted to life in Asia sometimes my world view and my boundaries between "medicine" and "life" are seriously challenged. The health project is staffed by young, locally trained nurses with a few older medical assistants trained at border camps. As our patients are generally managed well, I tend to assume a greater basic knowledge of anatomy and other medical sciences than may be the case.

I recently decided—having reached an impasse in my explanations of the relation between blood pressure in heart, lung, and liver—that it would be helpful to have an anatomy lesson with a fresh specimen. Mrs Vee, our office cleaner (and midday cook), was therefore sent to purchase the relevant pieces of pig. She returned with an impressive proportion of the internal organs—oesophagus, trachea, thyroid, heart, lungs, liver, and a piece of diaphragm. I particularly pointed out the heart and its resemblance to the human form (the aortic valve can be directly transplanted).

Part way through the lesson, I became aware of a most offensive smell (to my nose). I assumed that the specimen must be decomposing in the heat, but I was reassured that it was just the smell of "prahok"—fermented fish paste, a staple of the local diet—wafting in from the kitchen.

We continued the lesson until well after midday. No one seemed anxious to go home, but I gradually realised that, rather than still being enthralled with my teaching, the class was waiting for something. "It is lunch time," someone said.

"OK, we can go." said I and, pointing to the specimen, asked, "Where shall we dispose of this?"

A pitying look was exchanged—how stupid could she be? "Mrs Vee will cook it," Sok gently explained.

I was horrified. "But it has been sitting around all morning," I protested.

"Meat does that in the market," Sok pointed out.

"But we have been touching it."

"We have been wearing gloves." Sok replied, and then continued: "Oh, do you think we should give it to the poor people?"

"No," I said horrified, "if anyone should eat it we must."

So slices of the liver and heart were fried with ginger—and were, I have to admit, delicious. The slices of liver boiled and served with "prahok" were beyond my ability to eat. What happened to the lung was never clear, but I'm sure it nourished someone that evening.

Why was I so shocked? I am not a vegetarian and, having been raised on a farm, grew up eating offal. Was it the link with the human anatomy lessons of my youth that made me suddenly feel a kinship with this animal? Or is it that my medicine and "real life" are so far removed that I am uncomfortable when the gap closes? Perhaps the explanation is simpler: being well nourished and relatively affluent, I can afford the luxury of being particular about what I will and won't eat.

Janet Cornwall *project adviser, Servants Cambodia, Phnom Penh, Cambodia (janetmekong@yahoo.co.nz)*

We welcome articles up to 600 words on topics such as *A memorable patient, A paper that changed my practice, My most unfortunate mistake,* or any other piece conveying instruction, pathos, or humour. Please submit the article on http://submit.bmj.com Permission is needed from the patient or a relative if an identifiable patient is referred to. We also welcome contributions for "Endpieces," consisting of quotations of up to 80 words (but most are considerably shorter) from any source, ancient or modern, which have appealed to the reader.